About the Data

This dataset is from NASA (accessible here). It contains information about Near-Earth Comets (NECs, perihelion distance < 1.3 au*) and Asteroids (NEAs, perihelion distance < 1.3 au*) discovered recently. The data contains 202 rows of unique Near-Earth Objects (NEOs, inclusive of both asteroids and comets). Each record has:

- **Designation:** The name/number of the NEO
- **Discovery Date:** The date the NEO was first discovered
- **H (mag)/Magnitude:** Absolute magnitude, which is the intrinsic brightness of a celestial body, measured in magnitudes, computed as if viewed from a distance of 10 parsecs, or 32.6 light years
- **MOID (AU):** Minimum Orbit Intersection Distance, defined as the distance between the closest points of the orbits of two bodies, in this case the Earth and the NEO, measured in Astronomical Units (au)*
- **q (AU)/Perihelion:** Perihelion distance, which is the point in the orbit of an object where it comes closest to the sun, measured in Astronomical Units (au)*
- **Q (AU)/Aphelion:** Aphelion distance, which is the point in the orbit of an object where it is farthest from the sun, measured in Astronomical Units (au)*
- **period (yr)/Period:** The sidereal orbital period of the NEO, or the amount of time it takes the object to complete one orbit around the Sun, measured in Earth years
- **i (deg)/Inclination:** The angle of the object's orbit, with respect to the x-y ecliptic plane (what we would consider "flat"), measured in degrees
- **PHA:** Binary classification of whether or not the object is a Potentially Hazardous Asteroid (PHA), defined as NEAs with MOID<=0.05 au* and absolute magnitude (H)<=22.0
- **Orbit Class:** The object's classification based on its orbital characteristics. NEAs are divided into groups (Atira, Aten, Apollo and Amor) according to their perihelion distance (q), aphelion distance (Q) and their semi-major axes (a). NECs are classified as Encke-type Comet, Jupiter-family Comet, Halley-type Comet, or Comet, based on their orbital period and inclination.
    - **Atira:** NEAs whose orbits are contained entirely with the orbit of the Earth; a<1.0 au, Q<0.983 au
    - **Aten:** Earth-crossing NEAs with semi-major axes smaller than Earth's; a<1.0 au, Q>0.983 au
    - **Apollo:** Earth-crossing NEAs with semi-major axes larger than Earth's; a>1.0 au, q<1.017 au
    - **Amor:** Earth-approaching NEAs with orbits exterior to Earth's but interior to Mars'; a>1.0 au, 1.017<q<1.3 au
    - **Encke-type Comet:** Short-period comets (<200 year orbit) with orbits that do not reach the orbit of Jupiter
    - **Jupiter-family Comet:** Short-period comets with orbital periods less than 20 years and low inclinations (up to 30 degrees)
    - **Halley-type Comet:** Short-period comets with orbital periods of between 20 and 200 years and inclinations extending from 0 to more than 90 degrees

- ○ **Comet:** Long-period comets with an orbital period of 200+ years

* The Astronomical Unit (au) is a unit of length defined to be exactly equal to 149,597,870,700 meters, approximately the average distance between the Sun and Earth.

Problem Statement: We want to be able to understand the patterns in the NEO data, so we can predict potential hazards and be better equipped to identify additional NEOs in the future.
Business Questions:

- Were there certain years where more PHAs were discovered?
- Is there a relationship between Aphelion distance and Perihelion distance (in other words, are orbits roughly similar in shape or proportion regardless of scale)?
- What proportion of NEOs discovered recently are PHAs?
- What is the range of orbital periods?
- Are certain classes of NEOs more likely to be PHAs?

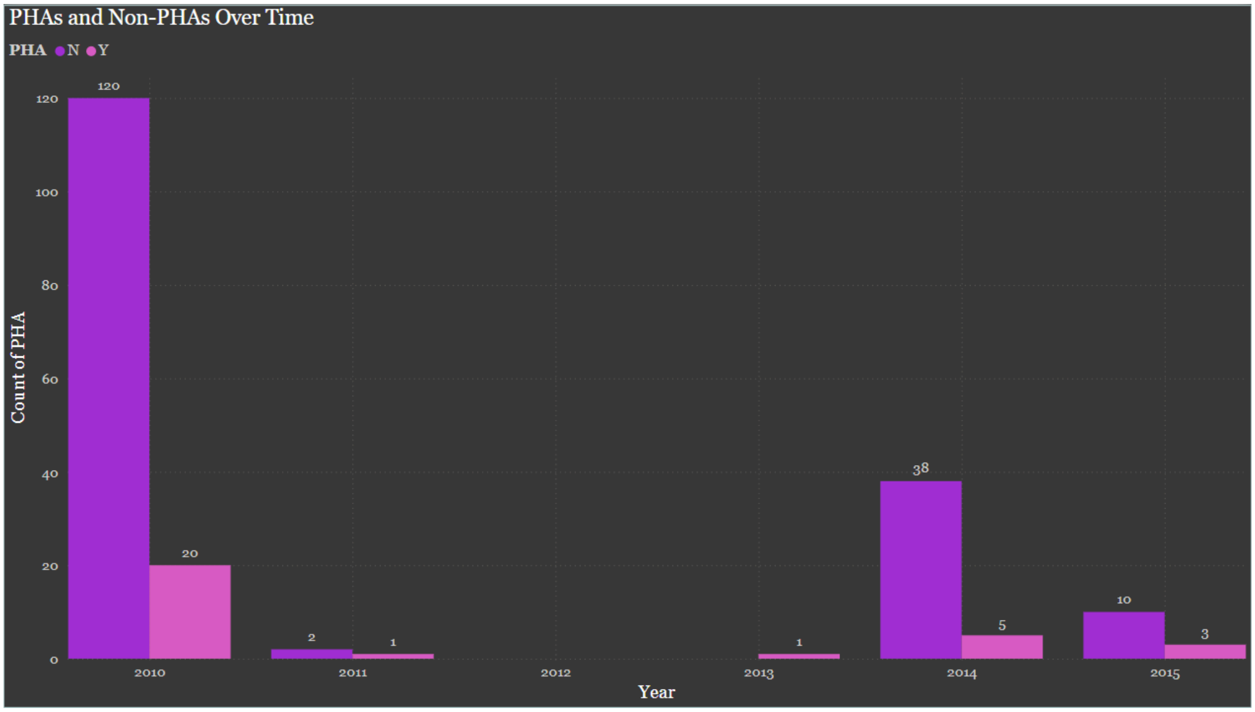
Data Cleaning

I started by checking for duplicates and missing data. There were no duplicates, but there were two records missing values for both Aphelion (Q) and Period. Because there was no way to calculate or discern these values and it was only two records out of 202, I dropped them. After that, there were 19 records missing both Magnitude (H) and PHA. To be classified as a PHA, an object must have MOID <= 0.05 au *and* Magnitude <= 22.0. The Magnitude value was missing for each of the records, but they all had MOID > 0.05, so they could all be classified as *not* PHAs (filled in as ‘N’). They were also all comets, which further disqualified them from classification as Potentially Hazardous *Asteroids*. Because all of the records missing Magnitude were comets and there were no other records of comets that had a listed Magnitude, there was insufficient information to extrapolate a meaningful value to impute, so I left those values as null. I did check for outliers, but given that this is an astronomical dataset, outliers are to be expected and generally should not be altered or discounted, so I left them alone. There is significant right skew in most of the numerical columns because of this.

To further optimize my data, I renamed the columns for clarity and reclassified the Discovery Date column as a datetime and the Orbit Class column as a categorical type.

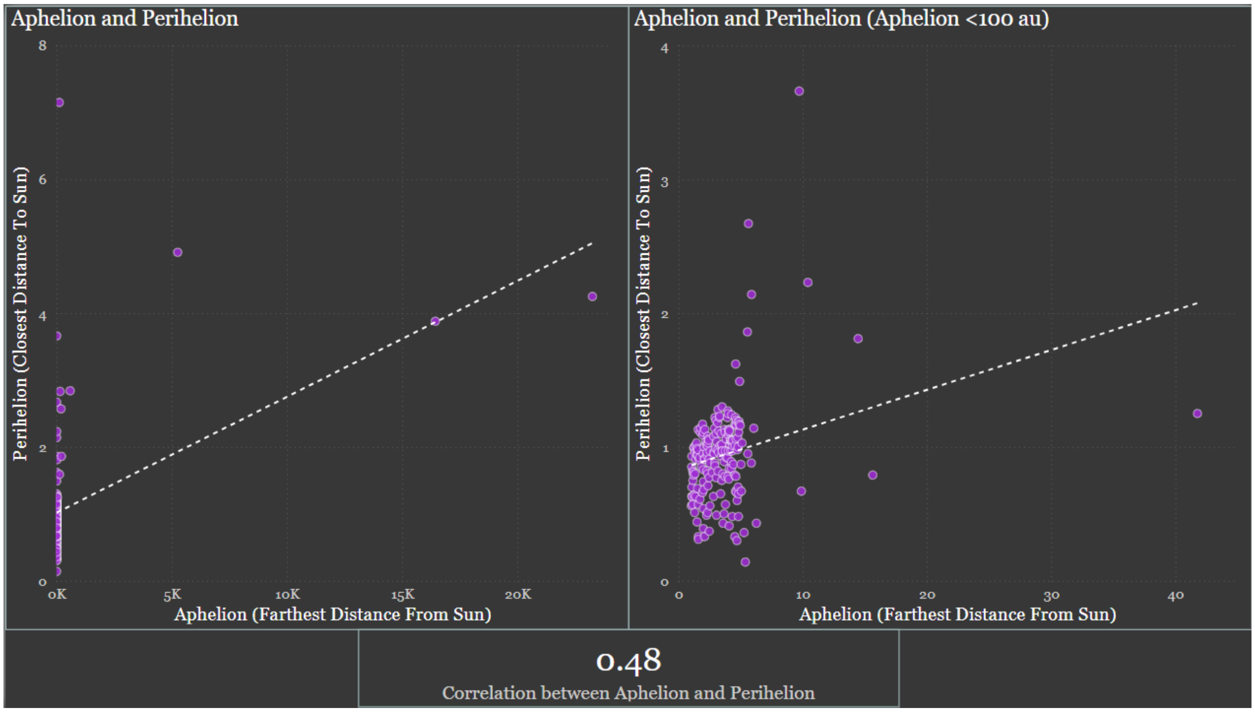
Analysis

Were there certain years where more PHAs were discovered?



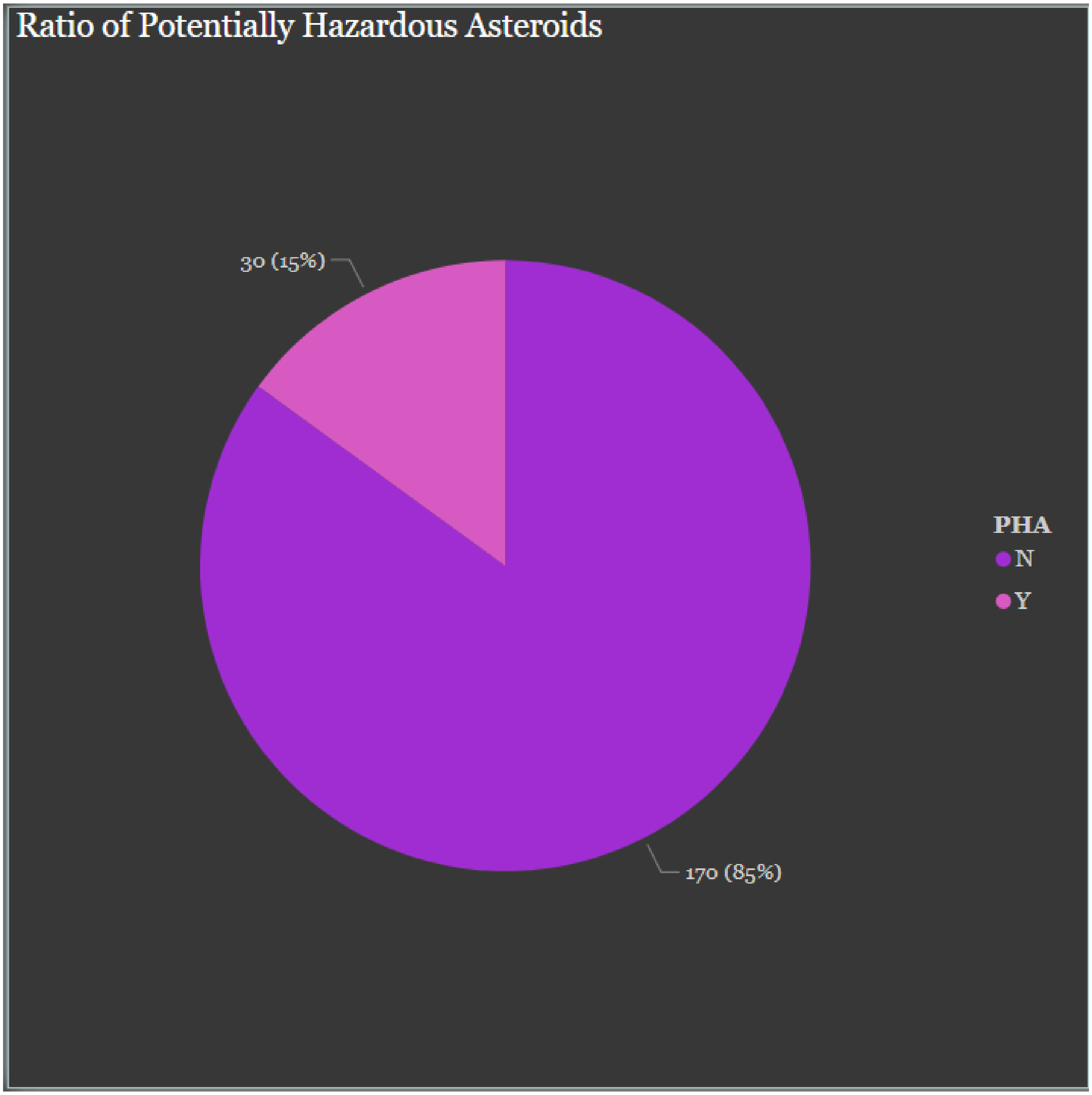
Of the 200 objects recorded in the dataset, 140 were discovered in 2010, 14.3% of which were classified as Potentially Hazardous Asteroids. Every year that NEOs were discovered, there was at least one PHA, and in 2013, the only object discovered was a PHA. Aside from 2013, each other year more non-PHAs were discovered, in line with the fact that there are far more non-PHAs than PHAs (which I will explore in more depth later in the analysis). The proportion of discoveries that are PHAs was highest in 2013 (100%), but the highest number found was in 2010 (20 PHAs). There is not enough information to indicate why more were found in certain years, so it may be worth examining technological advancements or specific research focuses around these years.

Is there a relationship between Aphelion distance and Perihelion distance?



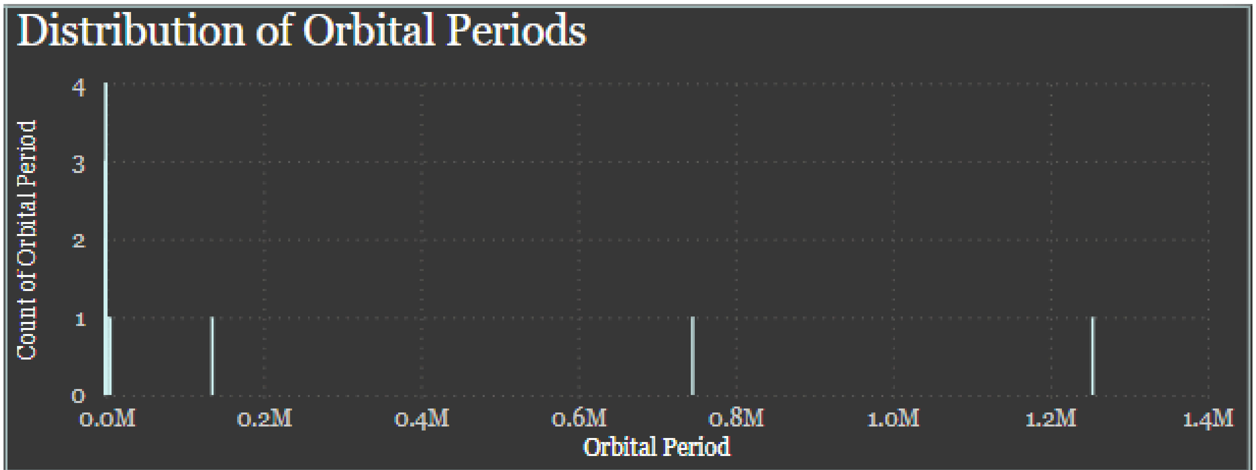
When examining the relationship between an NEO's Aphelion distance and Perihelion distance, I found that some outliers which travel extremely far from the Sun made it hard to visualize the data in a coherent way, so I created a second graph that zoomed in more on the majority of objects that stayed within 100 au of the Sun. An R2 value of 0.48 shows that there is a moderate positive correlation between Aphelion and Perihelion distances, which indicates some general similarity in orbit shapes and proportions, but not any true consistency. There are a fair amount of objects with more circular orbits (Aphelion and Perihelion are similar), but there are also a fair amount with widely varying proportions, including one NEO that gets as close as 4.25 au from the Sun yet also travels 23,355.11 au away at its farthest point, which gives it a *very* stretched elliptical orbit. The correlation is not strong enough to accurately calculate any missing data, but may be used in conjunction with other measures to estimate additional information about an NEOs orbit.

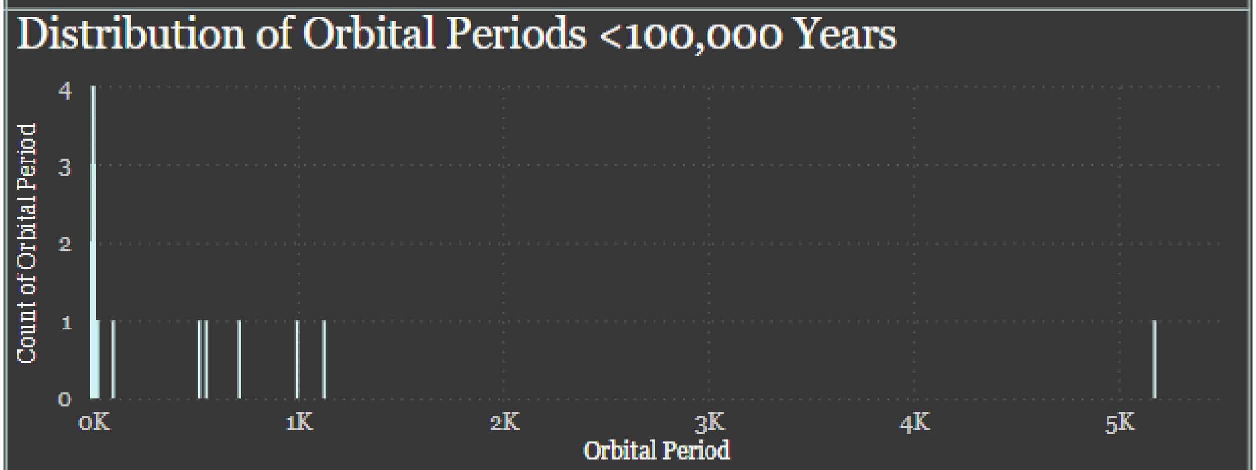
What proportion of NEOs discovered recently are PHAs?

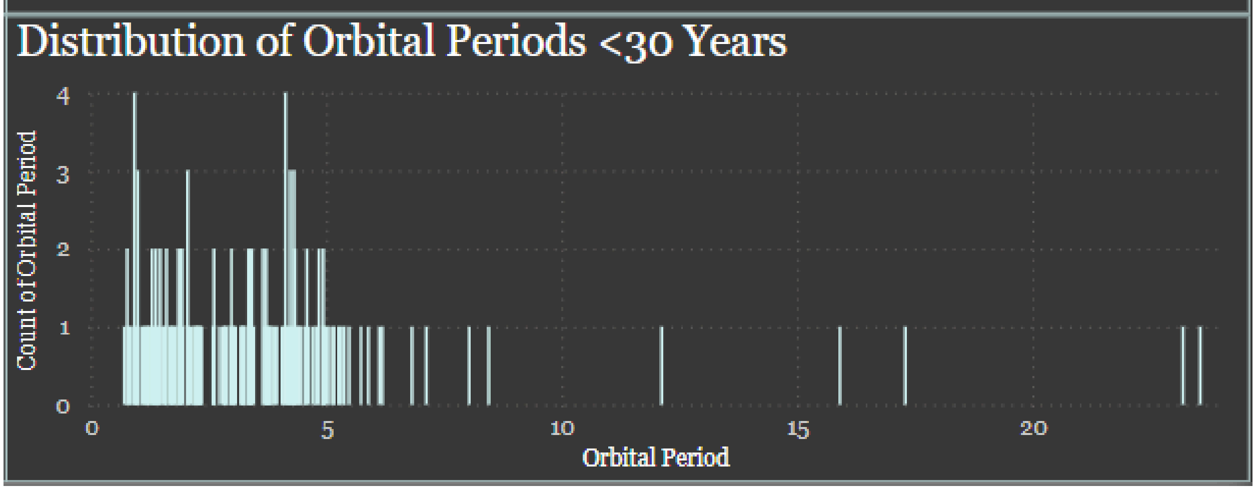


Of the Near Earth Objects in this dataset, discovered between 2010 and 2015, the vast majority (85%) were NOT Potentially Hazardous Asteroids, which is good news for us. NASA's classification system of PHAs can allow them to focus their budget and effort towards the 15% of NEOs that might be dangerous to us, rather than wasting time on the rest that won't interfere with our life on Earth.

What is the range of orbital periods?

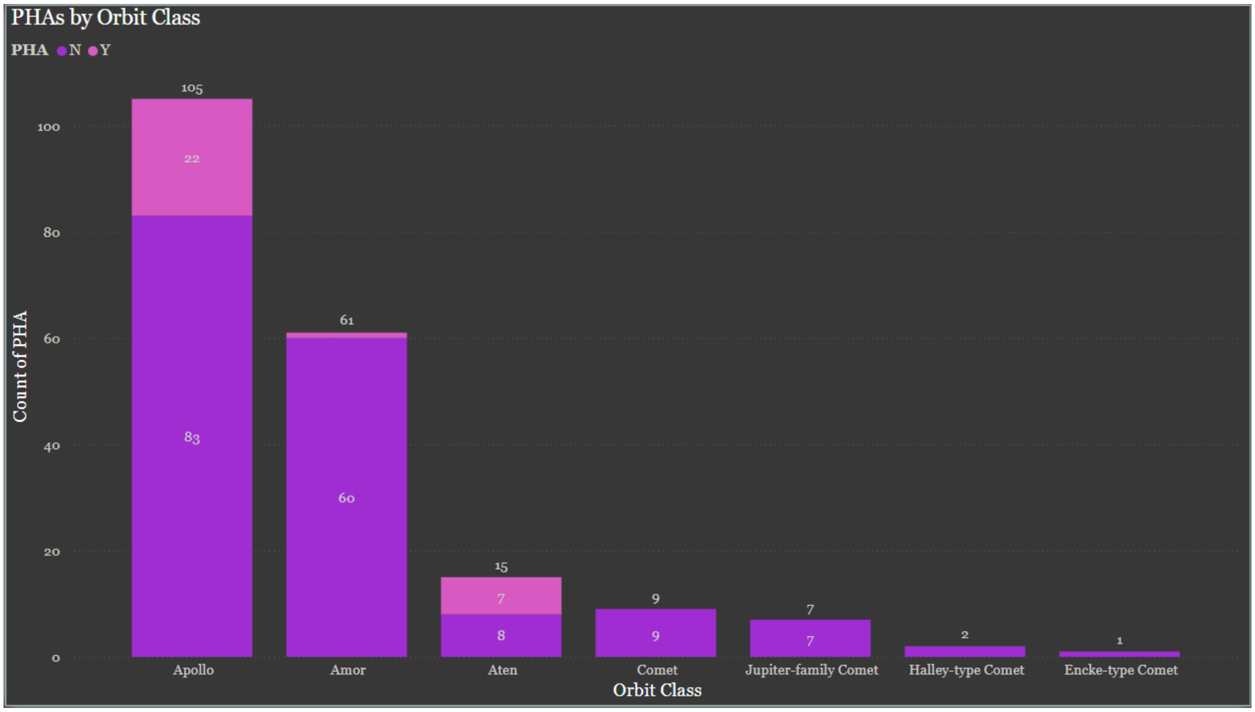






There is an enormous range in the length of orbital periods for NEOs. Although the vast majority are 5 Earth years or less, the durations span from 0.72 years all the way to 1,254,179.62 years (over 1.25 *million* years). With orbital periods that can far exceed a human lifespan, it is probably a good idea to continue to search for new NEOs regularly so we can be aware of ones that might only be close enough to see every couple thousand years or so.

Are certain classes of NEOs more likely to be PHAs?



Of the eight possible classifications in this dataset, four are types of comets, none of which are Potentially Hazardous Asteroids and there are no objects with the Atira classification (orbit entirely within Earth's orbit), so I'll just be examining the Apollo, Amor, and Aten NEOs. There are more Apollos that are PHAs than any other classification, but there are also just more Apollos overall, as they make up more than half of the recorded NEOs in the dataset. Atens have the highest proportion of PHAs (46.7%), but there are fewer overall. Only one of the 61 Amors is a PHA, which makes sense because Apollos and Atens cross Earth's orbit, while Amors do not, so Amors are far less likely to become a potential hazard. Based on this information, NASA would probably do well to focus more on tracking Apollos and Atens to optimize their resource usage to identify potential dangers to the Earth.

Conclusion

After analyzing the data, I found that approximately 15% of recently discovered NEOs are PHAs and the vast majority of these NEOs were discovered in 2010, though at least one PHA was discovered each year except 2012. Apollo and Aten asteroids are the most likely to be PHAs, as their orbits cross the Earth's orbit. NEOs have an enormous range in orbital periods, from 0.72 years to 1.25 million years, and there is a moderate strength relationship between the Perihelion and Aphelion of these orbits. This information can help NASA figure out where to focus risk prevention efforts and how to identify additional NEOs. The data in this project could be explored further by breaking down analyses by Orbit Class or other metrics and could potentially be expanded to create a model to classify NEOs or predict any missing data.